基因组学数据分析 第二次作业

目录

RNA-seq基本流程

软件安装

由于使用WSL/Debian环境，故在本地直接使用`apt`(和`pip`)安装。

```
apt-get install fastqc samtools bowtie2 python3-htseq
```

其中由于`htseq`检索结果为python包，故也可使用`pip`工具进行安装

```
pip3 install htseq
```

质量评估

使用`bash`命令行工具`fastqc`来对所有测序结果做质控

```
fastqc ./homework2_data/*.fastq
```

其中`SRR1039512_1.fastq`的部分结果如图

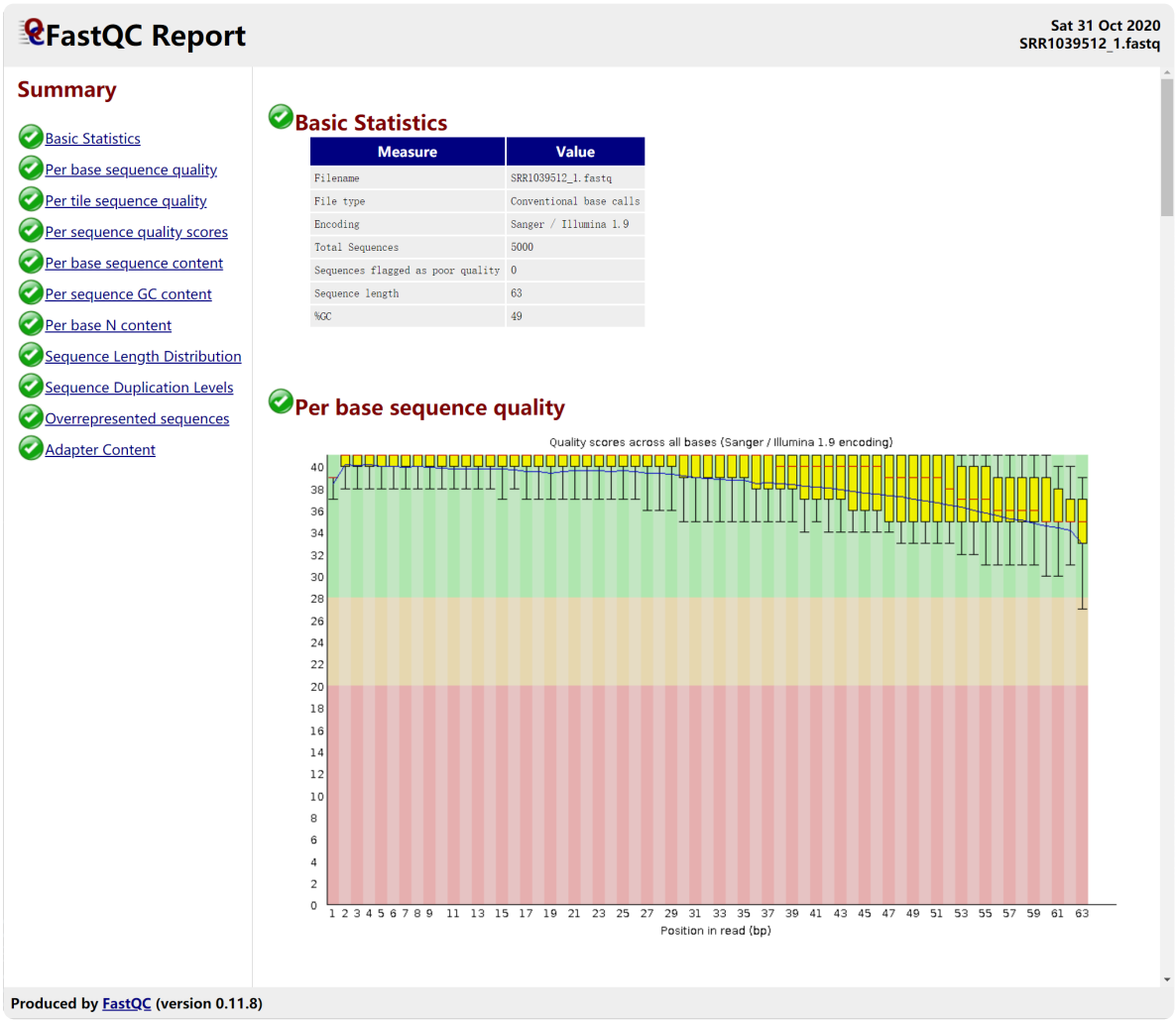


fig:sample result from fastQC

索引与比对

构建索引

使用`bowtie2`随带的`bowtie2-build`工具对`genome.fa`进行索引构建

```
bowtie2-build ./homework2_data/genome.fa ./homework2_data/genome.ref
```

> `bowtie2-build`命令行工具的使用帮助
>
> ```
> Bowtie 2 version 2.3.4.3 by Ben Langmead (langmea@cs.jhu.ed
> u, www.cs.jhu.edu/~langmea)
> Usage: bowtie2-build [options]* <reference_in> <bt2_index_ba
> se>
>     reference_in            comma-separated list of files wi
> th ref sequences
>     bt2_index_base          write bt2 data to files with thi
> s dir/basename
> ```

序列比对

使用`bowtie2`工具对双端测序结果进行比对

```
bowtie2 -x genome.ref -1 SRR1039508_1.fastq -2 SRR1039508_2.fastq -S SRR1
bowtie2 -x genome.ref -1 SRR1039509_1.fastq -2 SRR1039509_2.fastq -S SRR1
bowtie2 -x genome.ref -1 SRR1039512_1.fastq -2 SRR1039512_2.fastq -S SRR1
bowtie2 -x genome.ref -1 SRR1039513_1.fastq -2 SRR1039513_2.fastq -S SRR1
```

> `bowtie2`工具的使用帮助
>
> ```
> Bowtie 2 version 2.3.4.3 by Ben Langmead (langmea@cs.jhu.ed
> u, www.cs.jhu.edu/~langmea)
> Usage:
>   bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r>
>  | --interleaved <i>} [-S <sam>]
>
>   <bt2-idx>  Index filename prefix (minus trailing .X.bt2).
>              NOTE: Bowtie 1 and Bowtie 2 indexes are not com
> patible.
>   <m1>       Files with #1 mates, paired with files in <m2>.
>              Could be gzip'ed (extension: .gz) or bzip2'ed
>  (extension: .bz2).
>   <m2>       Files with #2 mates, paired with files in <m1>.
>              Could be gzip'ed (extension: .gz) or bzip2'ed
>  (extension: .bz2).
>   <r>        Files with unpaired reads.
>              Could be gzip'ed (extension: .gz) or bzip2'ed
>  (extension: .bz2).
>   <i>        Files with interleaved paired-end FASTQ reads
>              Could be gzip'ed (extension: .gz) or bzip2'ed
>  (extension: .bz2).
>   <sam>      File for SAM output (default: stdout)
>
>   <m1>, <m2>, <r> can be comma-separated lists (no whitespac
> e) and can be
>   specified many times.  E.g. '-U file1.fq,file2.fq -U file
> 3.fq'.
> ```
>
> `-x`之后为索引文件，不需要包含数字与后缀，与前一步`bowtie2-build`中命名保持一致
>
> `-1`和`-2`之后分别跟双端测序结果，可以写作逗号分隔列表或重复使用`-1`和`-2`参数，但是`-S <sam>`只会输出一个结果

序列比对结果

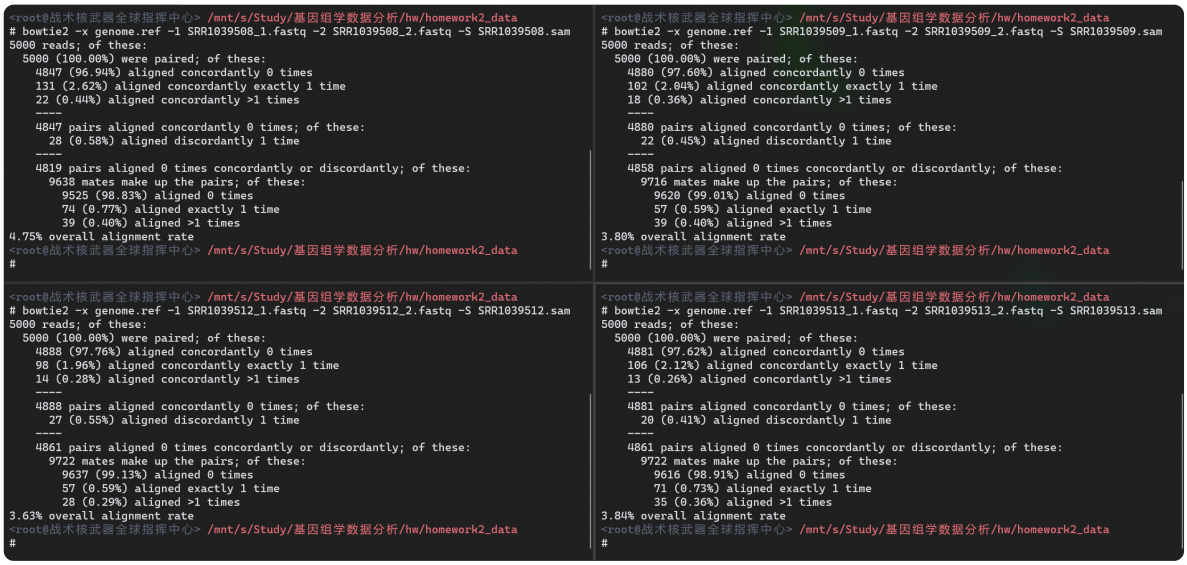


fig:bowtie2-res

基因计数

这里偷了个懒，写了个脚本跑一下命令行

```
import os
fileList = [x[:-5] for x in os.popen('ls homework2_data/*.sam').readlines
for eachFile in fileList:
    os.system('samtools view -Sb {}.sam > {}.bam'.format(eachFile,eachFil
    os.system('samtools sort -O bam -o {}.sorted.sam {}.bam'.format(eachF
    os.system('htseq-count -f bam -t exon -i gene_name {}.sorted.bam gene
```

> 1. `-Sb`代表输入格式为`.sam`（默认为`.bam`），输出格式为`.bam`（默认为`.sam`）
> 2. `-O bam`指定输出格式为`.bam`
>
>   `-o <output_file>`指定输出文件路径与文件名
> 3. `-f bam`指定输入文件格式为`bam`
>
>   `-t exon`指定只显示外显子的计数
>
>   `-i gene_name`指定以基因名而非ID作为显示
>
>   `2>*_htseq.log`将调试信息记录到日志文件中
>
>   管道符后`grep -v [[:space:]]0`将结果中计数为0的行全部排出，只留下有计数的行

计数汇总

接着偷懒，用python的字典类型完成对各个基因的计数

```
import os
fileList = [x[:-1] for x in os.popen('ls homework2_data/*_count.txt').rea
fileList
count={}
for eachFile in fileList:
    with open(eachFile,'r') as fileContent:
        for line in fileContent.readlines():
            if line[:2]=='__':
                continue
            line = line[:-1]
            res = line.split()
            count[res[0]]=count.get(res[0],0)+int(res[1])
gene_name = list(count.keys())
gene_count = list(count.values())
with open('homework2_data/count_tot.txt','w') as output:
    for i in range(len(gene_name)):
        output.write(gene_name[i]+'\t'+str(gene_count[i])+'\n')
```

计数结果

这里就不放截图了，少一点图片，让pdf稍微小一些

```
grep -e [[:space:]][[:digit:]][[:digit:]] homework2_data/count_tot.txt
```

```
FBLN1   66
MYH9    11
RPL3    27
TIMP3   25
LGALS1  14
```

R语言数据分析

差异表达分析

读入数据

```
raw_data <- read.csv("./homework2_data/raw_count.csv")
gene_count <- raw_data[, 2:5]
row.names(gene_count) <- raw_data[, 1]
```

`DESeq2` 差异分析

导入依赖包`DESEQ2`

```
library(DESeq2)
```

样品分组

```
sample_group <- factor(
    c("trt", "trt", "untrt", "untrt"),
    levels = c("trt", "untrt")
)
col_info <- data.frame(row.names = colnames(gene_count), sample_group)
```

差异矩阵

```
deseq_res <- DESeqDataSetFromMatrix(
    countData = gene_count,
    colData = col_info,
    design = ~sample_group
)
deseq_res_filtered <- deseq_res[rowSums(counts(deseq_res)) > 1, ]
deseq_output <- DESeq(deseq_res_filtered)
output_res <- results(deseq_output)
```

查看分析结果

```
summary(output_res)
```

```
out of 26003 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)       : 550, 2.1%
LFC < 0 (down)     : 705, 2.7%
outliers [1]       : 0, 0%
low counts [2]     : 9075, 35%
(mean count < 11)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

保存分析结果

```
output_res <- output_res[order(output_res$padj), ]
signif_diff_gene <- subset(output_res, padj < 0.01 & (log2FoldChange > 1
write.csv(signif_diff_gene, file = "./homework2_data/DESeq_trt-untrt.csv"
```

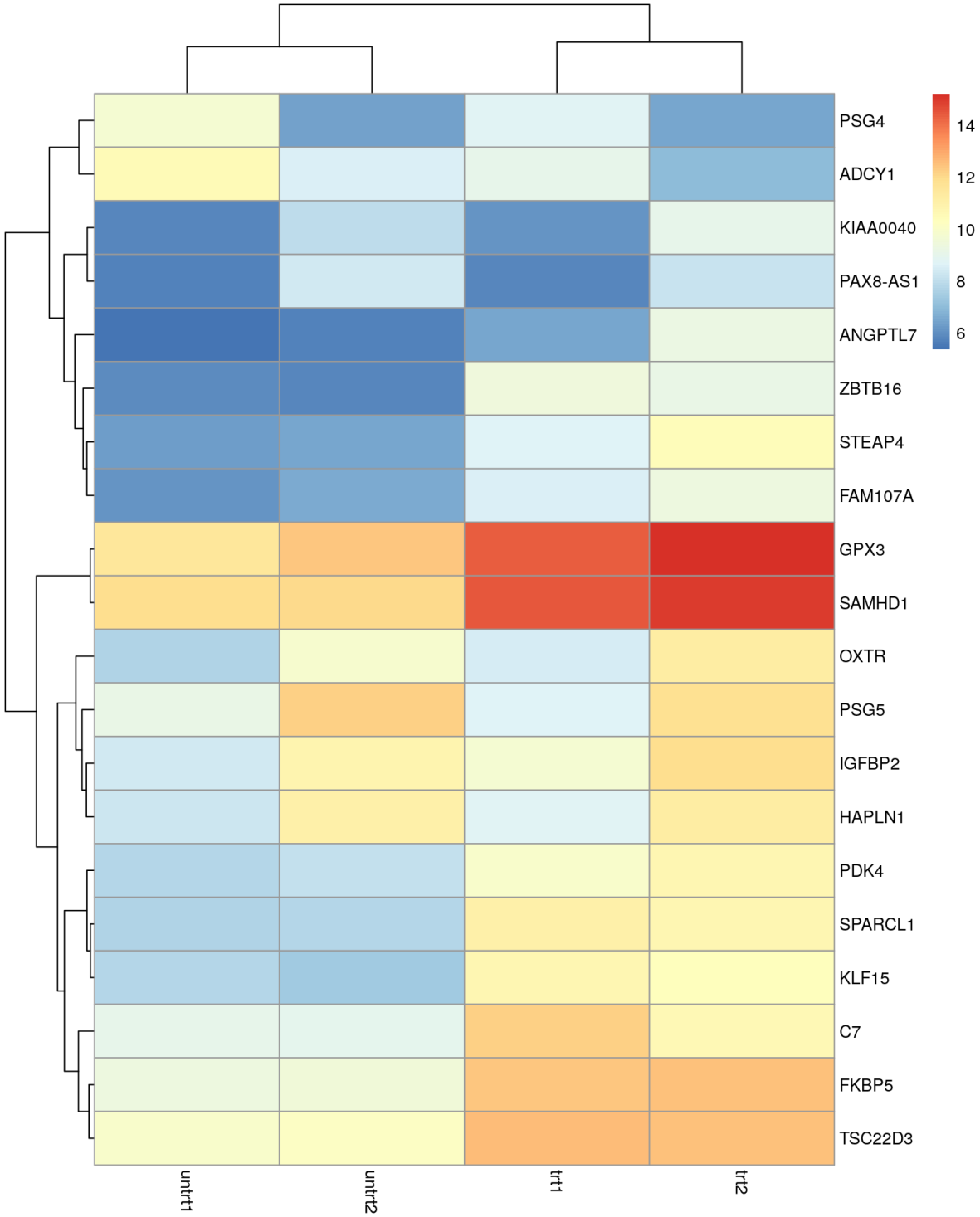
结果热图

导入依赖包

```
library(genefilter)
library(pheatmap)
```

热图绘制

```
deseq_output_uni <- rlogTransformation(deseq_output, blind = FALSE)
top_20_var_genes <- head(
    order(
        rowVars(assay(deseq_output_uni)),
        decreasing = TRUE
    ),
    20
)
res_mat <- assay(deseq_output_uni)[top_20_var_genes, ]
pheatmap(res_mat)
```



GO富集分析

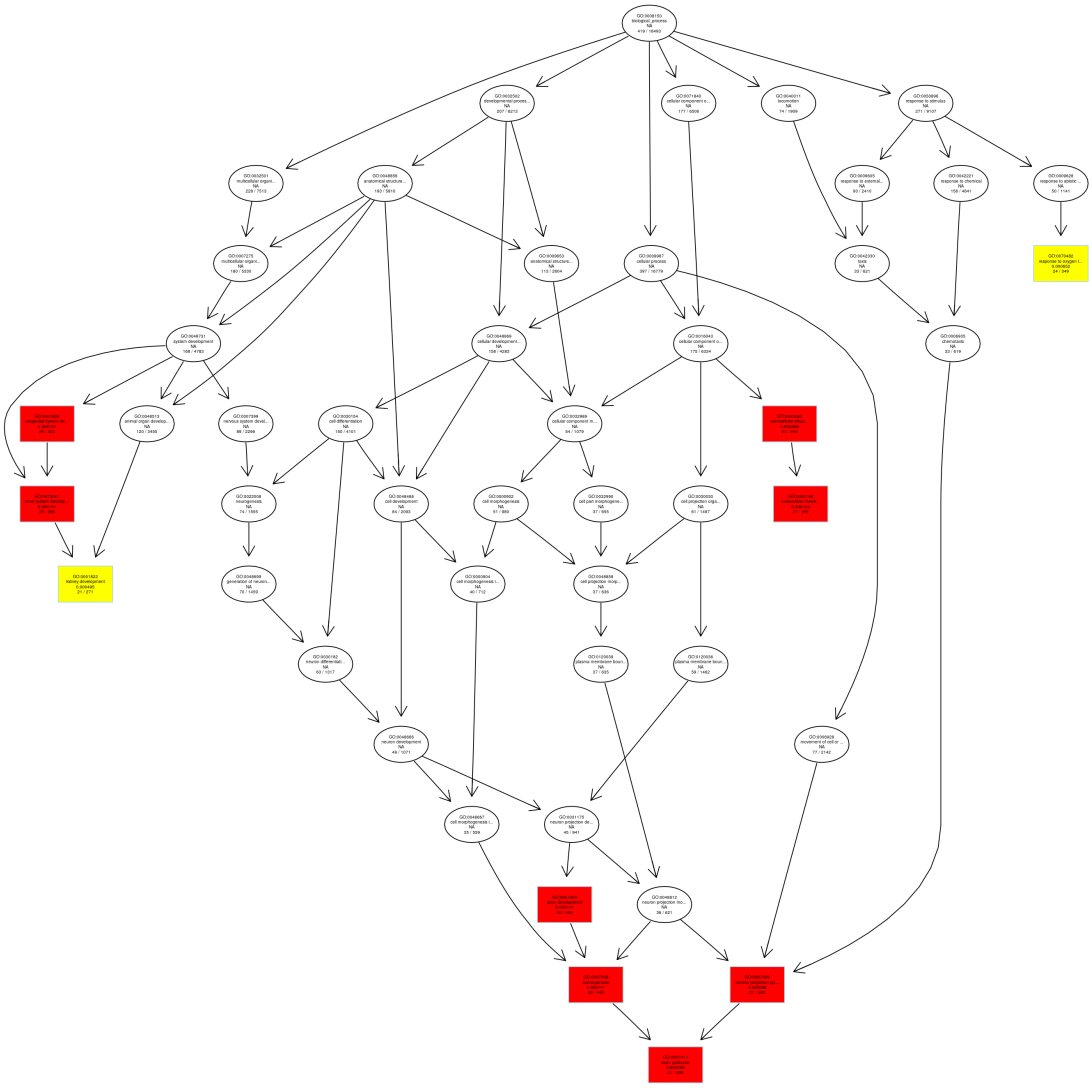
导入依赖包

```
library(clusterProfiler)
library(org.Hs.eg.db)
```

富集分析与作图

```
ego <- enrichGO(
    gene = rownames(signif_diff_gene),
    OrgDb = org.Hs.eg.db,
    keyType = "SYMBOL",
    ont = "BP",
    pAdjustMethod = "BH",
    pvalueCutoff = 0.01,
    qvalueCutoff = 0.05
)
plotGOgraph(ego)
```

```
groupGOTerms:   GOBPTerm, GOMFTerm, GOCCTerm environments built.


Building most specific GOs .....

    ( 5740 GO terms found. )


Build GO DAG topology ..........

    ( 5744 GO terms and 13185 relations. )


Annotating nodes ...............

    ( 18493 genes annotated to the GO terms. )
```



```
$dag
A graphNEL graph with directed edges
Number of Nodes = 48
Number of Edges = 75

$complete.dag
[1] "A graph with 48 nodes."
```